Research Article

Information Processing Features Can Detect Behavioral Regimes of Dynamical Systems

Rick Quax ¹, **Gregor Chliamovitch**,² **Alexandre Dupuis**,² **Jean-Luc Falcone**,² **Bastien Chopard**,² **Alfons G. Hoekstra**,^{1,3} and **Peter M. A. Sloom**^{1,3,4}

¹Computational Science Lab, University of Amsterdam, Amsterdam, Netherlands

²Department of Computer Science, University of Geneva, Geneva, Switzerland

³ITMO University, Saint Petersburg, Russia

⁴Complexity Institute, Nanyang Technological University, Singapore

Correspondence should be addressed to Rick Quax; r.quax@uva.nl

Received 11 September 2017; Revised 24 December 2017; Accepted 7 February 2018; Published 16 April 2018

Academic Editor: Dimitri Volchenkov

Copyright © 2018 Rick Quax et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In dynamical systems, local interactions between dynamical units generate correlations which are stored and transmitted throughout the system, generating the macroscopic behavior. However a framework to quantify exactly how these correlations are stored, transmitted, and combined at the microscopic scale is missing. Here we propose to characterize the notion of “information processing” based on all possible Shannon mutual information quantities between a future state and all possible sets of initial states. We apply it to the 256 elementary cellular automata (ECA), which are the simplest possible dynamical systems exhibiting behaviors ranging from simple to complex. Our main finding is that only a few information features are needed for full predictability of the systemic behavior and that the “information synergy” feature is always most predictive. Finally we apply the idea to foreign exchange (FX) and interest-rate swap (IRS) time-series data. We find an effective “slowing down” leading indicator in all three markets for the 2008 financial crisis when applied to the information features, as opposed to using the data itself directly. Our work suggests that the proposed characterization of the local information processing of units may be a promising direction for predicting emergent systemic behaviors.

1. Introduction

Emergent, complex behavior can arise from the interactions among (simple) dynamical units. An example is the brain whose complex behavior as a whole cannot be explained by the dynamics of a single neuron. In such a system, each dynamical unit receives input from other (upstream) units and then decides its next state, reflecting these correlated interactions. This new state is then used by (downstream) neighboring units to decide their new states and so on, eventually generating a macroscopic behavior with systemic correlations. A quantitative framework is missing to fully trace how correlations are stored, transmitted, and integrated, let alone to predict whether a given system of local interactions will eventually generate complex systemic behavior or not.

Our hypothesis is that Shannon’s information theory [1] can be used to construct, eventually, such a framework. In this viewpoint, a unit’s new state reflects its past interactions in the sense that it stores mutual information about the past states of upstream neighboring units. In the next time instant a downstream neighboring unit interacts with this state, implicitly transferring this information and integrating it together with other information into its new state and so on. In effect, each interaction among dynamical units is interpreted as a Shannon communication channel and we aim to trace the onward transmission and integration of information (synergy) through this network of “communication channels.”

In this paper we characterize the information in a single unit’s state at time t by enumerating its mutual information quantities with all possible sets of initial unit states ($t = 0$). We

generate initial unit states independently for the elementary cellular automata (ECA) application. Then we characterize “information processing” as the progression of a unit’s vector of information quantities over time (see Methods). The rationale behind this is as follows. The information in each initial unit state will be unique by construction, that is, have zero redundancy with all other initial unit states. Future unit states depend only on previous unit states and ultimately on the initial unit states (there are no outside forces). “Processing” refers, by our definition, to the fact that the initial (unique) pieces of information can be considered to disperse through the system in different directions and at different levels (synergy), while some of it dissipates and is lost. We can exactly trace all these directions and levels or every bit of information in the ECA due to the uniqueness of the initial information by construction. Therefore we would argue that we can then fully quantify the “information processing” of a system, implicitly, without knowing exactly which (physical) mechanism is actually responsible for this. We anticipate that this is a useful abstraction which will aid in distinguishing different emergent behaviors without being distracted by physical or mechanistic details. We first test whether this notion of information processing could be used to predict complex emergent behavior in the theoretical framework of ECA, under ideal conditions by construction. Next we also test if information processing could be used to detect a difference of systemic behavior in real financial time-series data, namely, the regimes before and after the 2008 crisis, despite the fact that obviously this data does not obey the strict ideal conditions.

The study of “information processing” in complex dynamical systems is a recently growing research topic. Although information theory has already been applied to dynamical systems such as elementary cellular automata, including, for instance, important work by Langton and Grassberger [2, 3], here we mean by “information processing” a more holistic perspective of capturing *all* forms of information simultaneously present in a system. As illustrative examples, Lizier et al. propose a framework to formulate dynamical systems in terms of distributed “local” computation: information storage, transfer, and modification [4] defined by individual terms of the Shannon mutual information sum (see (3)). For cellular automata they provide evidence for the long-held conjecture that so-called particle collisions are the primary mechanism for locally modifying information, and for a networked variant they show that a phase transition is characterized by the shifting balance of local information storage over transfer [5]. A crucial difference with our work is that we operate in the ensemble setting, as is usual for Shannon information theory, whereas Lizier et al. study a single realization of a dynamical system, for a particular initial state. (Although time-series data is strictly speaking a single realization, ensemble estimates are routinely made from such data by using sliding windows; see Methods.) Beer and Williams trace how task-relevant information flows through a minimally cognitive agent’s neurons and environment to ultimately be combined into a categorization decision [6] or sensorimotor behavior [7], using ensemble methods. Studying how local interactions lead to multiscale systemic

behavior is also a domain which benefits from information-theoretic approaches, such as those by Bar-Yam et al. [8, 9], Quax et al. [10, 11], and Lindgren [12]. Finally, extending information theory itself to deal with complexity, multiple authors are concerned with decomposing a single information quantity into multiple constituents, such as synergistic information, including James et al. [13], Williams and Beer [14], Olbrich et al. [15], Quax et al. [16], Chliamovitch et al. [17], and Griffith et al. [18, 19]. Although a general consensus on the definition of “information synergy” is thus still elusive, in this paper we circumvent this problem by focusing on the special case of independent input variables, in which case a closed-form formula (“whole-minus-sum”) is well-known and used.

2. Methods

2.1. Notational Conventions. Constants and functions are denoted by lower-case Roman letters. Stochastic variables are denoted by capital Roman letters. Feature vectors are denoted by Greek letters.

2.2. Model of Dynamical Systems. In general we consider discrete-time, discrete-state Markov dynamics. Let $X^t \equiv (X_1^t, X_2^t, \dots, X_N^t)$ denote the stochastic variable of the system state defined as the sequence of N unit states at time t . Each unit chooses its new state locally according to the conditional probability distribution $\Pr_i(X_i^{t+1} | X^t)$, encoding the microscopic system mechanics where i identifies the unit. The state space of each unit is equal and denoted by the set Σ . We assume that the number of units, the system mechanics, and the state space remain unchanged over time. Finally we assume that all unit states are initialized identically and independently (i.i.d.); that is, $\Pr(X^0) = \prod_{i=1}^N \Pr(X_i^0)$. The latter ensures that all correlations in future system states are generated by the interacting units and not an artifact of the initial conditions.

2.2.1. Elementary Cellular Automata. Specifically we focus on the set of 256 elementary cellular automata (ECA) which are the simplest discrete spatiotemporal dynamical systems possible [20]. Each unit has two possible states and chooses its next state deterministically using the same transition rule as all other cells. The next state of a cell deterministically depends only on its own previous state and that of its two nearest neighbors, forming a line network of interactions. That is,

$$\Pr_r(X_i^{t+1} | X^t) = \Pr_r(X_i^{t+1} | X_{i-1}^t, X_i^t, X_{i+1}^t). \quad (1)$$

There are 256 possible transition rules and they are numbered 0 through 255, denoted $r \in 0 \dots 255$. As initial state we take the fully random state so that no correlations exist already at $t = 0$; that is, $\Pr_r(X_i^{t=0}) = 1/2$ for all r and all i . The evolution of each cellular automaton is fully deterministic for a given rule, implying that the conditional probabilities in (1) can only be either 0 or 1. (This is nevertheless not a necessary condition in general.)

2.3. Quantifying the Information Processing in a Dynamical Model

2.3.1. Basics of Information Theory. We characterize each new unit state, determined probabilistically by $\Pr_i(X_i^{t+1} | X^t)$, by a sequence of Shannon communication channels, where each channel communicates information from a subset of X^t to X_i^{t+1} . In general, a communication channel $A \rightarrow B$ between two stochastic variables is defined by the one-way interaction $\Pr(B | A)$ and is characterized by the amount of information about the state A which transfers to the state B due to this interaction. The average amount of information stored in the sender's state A is determined by its marginal probability distribution $\Pr(A)$, which is known as its Shannon entropy:

$$H(A) = -\sum_a \Pr(A = a) \log_2 \Pr(A = a). \quad (2)$$

After a perfect, noiseless transmission, the information at the receiver B would share exactly $H(A)$ bits with the information stored at the sender A . After a failed transmission the receiver would share zero information with the sender, and for noisy transmission their mutual information is somewhere in between. This is quantified by the so-called mutual information:

$$\begin{aligned} I(A : B) &= H(A) - H(A | B) \\ &= \sum_{a,b} \Pr(A = a, B = b) \log_2 \frac{\Pr(A = a, B = b)}{\Pr(A = a) \Pr(B = b)}. \end{aligned} \quad (3)$$

The conditional variant $H(A | B)$ obeys the chain rule $H(A, B) = H(B) + H(A | B)$ and is written explicitly as

$$\begin{aligned} H(A | B) &= -\sum_b \Pr(B = b) \\ &\cdot \sum_a \Pr(A = a | B = b) \log_2 \Pr(A = a | B = b). \end{aligned} \quad (4)$$

This denotes the remaining entropy (uncertainty) of A given that the value for B is observed. For intuition it is easily verified that the case of statistical independence, that is, $P(B | A) = P(B)$, leads to $H(A | B) = H(A)$ which makes $I(A : B) = 0$, meaning that B contains zero information about A . At the other extreme, $B = A$ would make $H(A | B) = 0$ so that $I(A : B) = H(A)$, meaning that B contains the maximal amount of information needed to determine a unique value of A .

2.3.2. Characterizing the Information Stored in a Unit's State. First we characterize the information stored in a unit's state at time step t , denoted by $\vec{\phi}_i^t(r)$, as the ordered sequence of mutual information quantities with all possible sets of unit states at time $t = 0$; that is,

$$\vec{\phi}_i^t(r) \equiv (I_r(X_i^t : S) : S \in 2^{X^0}). \quad (5)$$

Here 2^{X^0} denotes the (ordered) power set notation for all subsets of stochastic variables of initial cell states. (Note

though that in practice not infinitely many initial cell states are needed; for instance, for an ECA at time t only the nearest $1 + 2t$ initial cell states are relevant.) We will refer to $\vec{\phi}_i^t(r)$ as the sequence of information features of unit i at time t . The subscript notation $I_r(\cdot : \cdot)$ implies that the rule-specific (conditional) probabilities $\Pr_r(X_i^t | X^0)$ and $\Pr_r(X_i^t)$ are used to compute the mutual information. We use the subscript i for generality to emphasize that this feature vector pertains to each single unit (cell) in the system, even though in the specific case of ECA this subscript could be dropped as all cells are indistinguishable.

In particular we highlight the following three types of information features. The “memory” of unit i at time t is defined as the feature $I_r(X_i^t : X_i^0) \in \vec{\phi}_i^t(r)$, that is, the amount of information that the unit retains about its own initial state. The “transfer” of information is defined as nonlocal mutual information such as $I(X_i^t : X_j^0) \in \vec{\phi}_i^t(r)$ ($i \neq j$). Nonlocal mutual information must be due to interactions because the initial states are independent (all pairs of units have zero mutual information). Finally we define the integration of information as “information synergy,” an active research topic in information theory [4, 14, 16, 19, 21–23]. The information synergy in X_i^t about X^0 is calculated here by the well-known whole-minus-sum (WMS) formula $I(X_i^t : X^0) - \sum_{j=1}^N I(X_i^t : X_j^0)$. The WMS measure directly implements the intuition of subtracting the information carried by individual variables from the total information. However the presence of correlations among the X_j^0 would be problematic for this measure, in which case it can become negative. In this paper we prevent this by ensuring that the X_j^0 are uncorrelated. In this case it fulfills various proposed axiomatizations for synergistic information known thus far, particularly PID [14, 15] and SRV [16].

Information synergy (or “synergy” for short) is not itself a member of $\vec{\phi}_i^t(r)$ but it is fully redundant given $\vec{\phi}_i^t(r)$ since each of its terms is in $\vec{\phi}_i^t(r)$. Therefore we will treat synergy features as separate single features in our results analysis while we do not add them to $\vec{\phi}_i^t(r)$.

2.4. Predicting the Class of Dynamical Behavior Using Information Processing Features. We require a classification of (long-term) dynamical behavior of the systems under scrutiny which is to be predicted by the information features. In this paper we choose the popular Wolfram classification for the special case of ECA.

2.4.1. Behavioral Class of a Rule. Wolfram observed empirically that each rule $r \in 0, \dots, 255$ tends to evolve from a random initial state to one of only four different classes of dynamical behavior [20]. These de facto established behavioral classes are

- (1) homogeneous (all cells end up in the same state);
- (2) periodic (a small cycle of repeating patterns);
- (3) chaotic (pseudorandom patterns);

- (4) complex (locally stable behavior and long-range interactions among patterns).

These classes are conventionally numbered 1 through 4, respectively. We obtained the class number for all 256 rules from Wolfram Alpha [24] and denote it by $c(r) \in 1, \dots, 4$. When the rule number is treated as a stochastic variable it will be denoted R ; similarly, if the class number is treated as stochastic variable it will be denoted $c(R)$.

2.4.2. Predictive Power of the Information Processing Features.

We are interested in the inference problem of predicting the class number $c(R)$ based on the observed information features $\vec{\phi}_i^t(R)$. Here the rule number R is considered a uniformly random stochastic variable, making in turn functions such as $c(R)$ also stochastic variables. We formalize the prediction problem by the conditional probabilities $\Pr(c(R) | \vec{\phi}_i^t(R))$. That is, given only the sequence of information features of a specific (but unknown) rule $r \in R$ at time t , what is the probability that the ECA will eventually exhibit behavior of class $c(r)$? We can interpret this problem as a communication channel $c(R) \rightarrow \vec{\phi}_i^t(R)$ and quantify the predictive power of $\vec{\phi}_i^t(R)$ using the mutual information $I(c(R) : \vec{\phi}_i^t(R))$. The predictive power is thus zero in case the information features do not reduce the uncertainty about $c(R)$, whereas it achieves its maximum value $H(c(R))$ in case a sequence of information features $\vec{\phi}_i^t(R)$ always uniquely identifies the behavioral class $c(R)$. We will normalize the predictive power as $I(c(R) : \vec{\phi}_i^t(R))/H(c(R))$. For the Wolfram classification, $H(c(R)) \approx 1.133$, which is lower than the maximum possible (2.0) since there are relatively many rules with class 2 behavior and not many complex rules.

Note that a normalized predictive power of, say, 0.75 does not necessarily mean that 75% of the rules can be correctly classified. Our definition yields merely a relative measure where 0 means zero predictive power, 1 means perfect prediction, and intermediate values are ordered such that a higher value implies that a more accurate classification algorithm could in principle be constructed. The benefit of our definition based on mutual information is that it does not depend on a specific classifier algorithm; that is, it is model-free. Indeed, the use of mutual information as a predictor of classification accuracy has become the de facto standard in machine learning applications [25, 26].

2.4.3. Selecting the Principal Features. Some information features are more predictive than others for determining the behavioral class of a rule. Therefore we perform a feature selection process at each time t to find these “principal features” as follows. First we extend the set of information features $\vec{\phi}_i^t(R)$ by the following set of synergy features:

$$\vec{\sigma}_i^t(r) \equiv \left(I_r(X_i^t : S) - \sum_{s_j \in S} I_r(X_i^t : S_j) : S \in 2^{X^{t=0}} \right). \quad (6)$$

Their concatenation makes the extended ordered feature set, now written in the form of stochastic variables:

$$\vec{\Phi}_i^t(R) \equiv \left(\vec{\phi}_i^t(R), \vec{\sigma}_i^t(R) \right). \quad (7)$$

The extended feature set $\vec{\Phi}_i^t(R)$ has no additional predictive power compared to $\vec{\phi}_i^t(R)$, so for any inference task $\vec{\Phi}_i^t(R)$ and $\vec{\phi}_i^t(R)$ are equivalent. That is, the synergy features $\vec{\sigma}_i^t(R)$ are completely redundant given $\vec{\phi}_i^t(R)$ since each of its terms is a member of $\vec{\phi}_i^t(R)$. The reason for adding them separately to form $\vec{\Phi}_i^t(R)$ is that they have a clear meaning as information which is stored in a collection of variables while not being stored in any individual variable. We are interested to see whether this phenomenon plays a significant role in generating dynamical behaviors.

We define the first principal feature $\Phi_{i,1}^t$ at time t as maximizing its individual predictive power, quantified by a mutual information term as explained before, as

$$\Phi_{i,1}^t \equiv \arg \max_{\Phi_R \in \vec{\Phi}_i^t(R)} I(\Phi_R : c(R)). \quad (8)$$

Here, again rule number R is treated as a uniformly random stochastic variable with $\Pr(R) = 1/256$ which in turn makes Φ_R and $c(R)$ stochastic variables. In words, $\Phi_{i,1}^t$ is the single most predictive information feature about the behavioral class that will eventually be generated. More generally, the principal set of n features is identified in similar spirit; namely,

$$\Phi_{i,n}^t \equiv \arg \max_{\vec{\Phi}_R \subseteq (\vec{\Phi}_i^t(R))^n} I(\vec{\Phi}_R : c(R)). \quad (9)$$

2.4.4. Information-Based Classification of Rules. The fact that Wolfram’s classification relies on the behavior exhibited by a particular initial configuration makes the complexity class of an automaton dependent on the initial condition. Moreover, there is no universal agreement regarding how “complexity” should be defined and various alternatives to Wolfram’s classification have been proposed, although Wolfram’s remains by far the most popular. Our hypothesis is that the complexity of a system has very much to do with the way it processes information. Therefore we also attempt to classify ECA rules using only their informational features.

We use a classification algorithm which takes as input the 256 vectors of information features and computes the Euclidean distance between these vectors. The two vectors nearest to each other are clustered together. Then the remaining nearest elements or clusters are clustered together. The distance between two clusters is defined as the distance between the most distant elements in each cluster. The result is a hierarchy of clusters with different distances which we visualize as a dendrogram.

2.5. Computing Information Processing Features in Foreign Exchange Time-Series. In the previous section we

define information processing features for the simplest (one-dimensional) model of discrete dynamical systems. In the second part of this paper we aim to investigate if information features can distinguish “critical” regimes in the real complex dynamical system of the foreign exchange market. Most importantly, we are interested in the behavior of the information features before, at, and after the start of the 2008 financial crisis, which is commonly taken to coincide with the bankruptcy of Lehman Brothers on September 15, 2008. We consider two types of time-series datasets in which the dynamical variables can be interpreted to form a one-dimensional system in order to stay as close as possible to the ECA modeling approach.

The information features can then be computed as discussed above, except that each mutual information term is now estimated directly from the data. This estimation is performed within a sliding window of length w up to time point t which enables us to see how the information measures evolve over time t . For instance, the memory of variable X at time t' will be measured as $I(X^{t'} : X^{t'-1})$ where the joint probability distribution $\Pr(X^{t'}, X^{t'-1})$ is estimated using only the data points $X^{t'-w}, \dots, X^{t'}$. Details regarding the estimation procedure are given in the following subsection. The i th time-series in a dataset will be denoted by subscript as in X_i^t .

2.5.1. Estimating Information Processing Features from the Data. The mutual information between two financial variables (time-series) at time t is estimated using the k -nearest-neighbor algorithm using the typical setting $k = 3$ [27]. This estimation is calculated using a sliding window of size w leading up to and including time point t , after first detrending each time-series using log-returns. For all results we will use 200 uniformly spaced values for t over the dataset, starting at datapoint w and ending at the length of the dataset. Thus windows partially overlap. Parameter w is evaluated for robustness in the Supplementary Materials (available here).

We calculate the “memory” (M) of a time-series i as $I(X_i^t : X_i^{t-1})$ and the average “transfer” (T) as $[I(X_i^t : X_{i-1}^{t-1}) + I(X_i^t : X_{i+1}^{t-1})]/2$. That is, whereas in the ECA model we calculated the mutual information quantities with respect to the initial state of the model, here we use consecutive time points, effectively treating X_i^{t-1} as the initial state and characterizing only the single time step to X_i^t .

For calculating the synergy measure (S) we apply a correction which makes this measure distributed around zero. The reason is that the WMS measure of (6) assumes independence among the stochastic (initial) state variables S_i , which for the real data are taken to be the previous day’s time-series values. When this assumption is violated, it can become strongly negative and, more importantly, cointegrated with the memory and transfer features whose sum will then dominate the synergy feature. We remedy this by rescaling the sum of the memory and transfer features which are subtracted in (6) to equal the average value of the total information (positive term in (6)). In formula,

a constant c is inserted into the WMS formula, leading to

$$I(X_i^t : X_{i-1}^{t-1}, X_i^{t-1}, X_{i+1}^{t-1}) - c \cdot \sum_{\delta \in \{-1, 0, +1\}} I(X_i^t : X_{i-\delta}^{t-1}) \quad (10)$$

for a given set of initial cell states S . c is fitted such that this WMS measure is on average 0 for all sliding windows over the dataset. This rejects the cointegration null-hypothesis between total information and the subtracted term at the 0.05 significance level ($p \approx 0.002$) in this dataset. This results in the synergy feature being distributed around zero and being independent of the sum of the other two features so that it may functionally be used as part of the feature space for feature selection; however, the value itself should not be trusted as quantifying precisely the notion synergy.

2.5.2. Description of the Foreign Exchange Data. The first data we consider are time-series of five foreign exchange (FX) daily closing rates (EUR/USD, USD/JPY, JPY/GBP, GBP/CHF, and CHF/EUR) for the period from January 1, 1999, to April 21, 2017 [28]. Each currency pair has a causal dependence on its direct neighbors in the order listed because they share a common currency. For instance, if the EUR/USD rate changes then USD/JPY will quickly adjust accordingly (among others) because the rate imbalance can be structurally exploited for profit. In turn, among others through the rate JPY/EUR (not observed in this dataset) the rate EUR/USD will also be adjusted due to profit-making imbalances, eventually leading to all neighboring rates returning to a balanced situation.

2.5.3. Description of the Interest-Rate Swap Data. The second data are interest-rate swap (IRS) daily rates for the EUR and USD market [11]. The data spans over twelve years: the EUR data from January 12, 1998, to August 12, 2011, and the USD data from April 29, 1999, to June 6, 2011. The datasets consist of 14 and 15 times to maturity (durations), respectively, ranging from 1 year to 30 years. Rates for nearby maturities have a dependency because the higher maturity can be constructed by the lower maturity plus a delayed (“forward”) short-term swap. This basic mechanism between maturities leads to generally monotonically upward “swap curves.”

3. Results

3.1. Predicting the Wolfram Class of ECA Rules Using Information Processing Features

3.1.1. Information Processing in the First Time Step. The information processing occurring in the first time step of each ECA rule $r \in 0 \dots 255$ is characterized by the corresponding feature set $\vec{\Phi}_i^{t=1}(r)$, consisting of 7 time-delayed mutual information quantities ($\vec{\phi}_i^{t=1}(R)$, (5)) and 4 synergy quantities ($\vec{\sigma}_i^{t=1}(R)$, (6)). We show three selected features (memory, total transfer, and total synergy) for all 256 rules as points in a vector-space in Figure 1 along with each rule’s Wolfram class as a color code. It is apparent that the three features already

FIGURE 1: Three selected features from $\vec{\Phi}_i^{t=1}(r)$ for each ECA rule $r \in 0 \dots 255$, namely, $I_r(X_i^{t=1} : X_i^{t=0})$ (“memory”), $I_r(X_i^{t=1} : X_{i-1}^{t=0}) + I(X_i^{t=1} : X_{i+1}^{t=0})$ (“transfer”), and $I_r(X_i^{t=1} : X_{i-1}^{t=0}, X_i^{t=0}, X_{i+1}^{t=0}) - \sum_{\delta \in \{-1, 0, +1\}} I_r(X_i^{t=1} : X_{i+\delta}^{t=0})$ (“synergy”). Each dot corresponds to a rule r and is color-coded by its Wolfram class $c(r)$, namely, black for the simple homogeneous (24) and periodic behaviors (196), green for complex behavior (10), and red for chaotic behavior (26). The transparency of a point indicates its distance away from the viewer, with more transparent points being farther away. A small random vector with average norm 0.02 is added to each point in order to make rules with equal information features still visible. The gray plus signs are the projections of the 3D points on the two visible side faces (the S-M face is occluded) for better visibility of the positions of the points.

partially separate the behavioral classes. Namely, it turns out that chaotic and complex rules tend to have high synergy, low information memory, and low information transfer. Figure 1 also relates intuitively to the classic categorization problem in machine learning; namely, perfect prediction would be equivalent to the existence of hyperplanes that perfectly separate all four behavior classes. In the case of ECA the information features are deterministic calculations for each rule number. Thus $\vec{\Phi}_i^{t=1}(r)$ forms a discrete distribution of 256 points such that separability implies that no two rule numbers fall on exactly the same point in this information space.

3.1.2. Predictive Power of Information Processing Features Over Time. The single most predictive information feature is *synergy*, as shown in Figure 2. Its predictive power is about 0.37 (where 1.0 would mean perfect prediction). The most predictive *pair* of features is formed by adding information *transfer* at 0.43, so adding the information transfer feature increases the predictive power by 0.06. The information transfer feature by itself has actually over three times this predictive power at 0.19, showing that two or more features can significantly overlap in their prediction of the behavioral class. The total predictive power of all information processing features at $t = 1$ is 0.49, formed by 4 of the 11 possible information features.

For the second time step (Figure 2) we again find that the most predictive information feature is synergy. An intriguing difference however is that it is now significantly more predictive at 0.90. This means that already at $t = 2$ there is a single

information characteristic of dynamical behavior (i.e., synergy), which explains the vast majority of the entropy $H(c(R))$ of the behavioral class that will eventually be exhibited. A second intriguing difference is that the maximum predictive power of 0.98 is now achieved using only 3 out of 57 possible information features, where 4 features were needed at $t = 1$.

Finally, for $t = 3$ we find that only 2 information features are needed to achieve the maximum possible predictive power of 1.0; that is, the values for these two features uniquely identify the behavior class. Firstly this confirms the apparent trend that fewer information features capture more of the relevant dynamical behavior as time t increases. Secondly we find again that synergy is the single most predictive feature. In addition, we find again that the best secondary feature is a peculiar combination of memory and the two longest-range transfers, as in $t = 2$. Including the intermediate transfers (so adding I1111111 instead of I1001001 as second feature) actually only slightly convolutes the prediction: adding them in $t = 2$ reduces predictive power by 0.028, whereas in $t = 3$ it does not reduce the predictive power at all. In $t = 1$ there are no intermediate transfers possible since there are only three predecessors of a cell's state, and apparently then it pays off to leave out memory (which would reduce power by 0.025 if added).

One could argue that the quick separation of the points in information space is hardly surprising because a high-dimensional space is used to separate only a small number (256) of discrete points. To validate that the predictive power

FIGURE 2: Predictive power of the optimal set of n information features as function of n (thick blue line with round markers). For each n the best set of information features is listed: S means “synergy” and I means mutual information, followed by a bitmask of which neighbor cell’s initial state is included in the feature (middle bit is the cell itself). For example, I010 indicates the “memory” feature $I(X_i^1 : X_i^0)$, and S111 denotes $I(X_i^{t=1} : X_i^{t=0}) - \sum_j I(X_i^{t=1} : X_j^{t=0})$. The thin black line with error bars is the “base line” predictability for n features, obtained by randomizing the pairing of 256 information features with their class identifier. The error bar indicates the 95% confidence interval of the distribution of predictive power under the null-hypothesis of zero correlation between information features and class identifier. Finally, the small black markers indicate the predictive powers of all other information feature sets.

values of the information features are indeed meaningful we also plot the expected “base line” prediction power in each subfigure in Figure 2 along with the 95% confidence interval. The base line is the null-hypothesis formed by randomizing the pairing of information feature values with class identifiers; that is, it shows the expected predictive power of having the same number and frequencies of feature values but sampled with zero correlation with the classification to make the separability meaningless. This results in a statistical test with a null-hypothesis of the information features and Wolfram classification being uncorrelated. We find that the predictive power of the information features is always significantly above the base distributions. Therefore we consider the meaningfulness (or “surprise”) of the information features’ separability validated; that is, we reject the null-hypothesis at the 95% confidence level that the observed quick separation in information space is meaningless and merely due to dimensionality.

3.1.3. Relation to Langton’s Parameter. Langton’s λ parameter [2] is the most well-known single feature of an ECA rule which partially predicts the Wolfram class. It is a single scalar computed for rule r as $\lambda_r = \Pr(X_i^1 = 0 \mid r)$. It is known that the λ parameter is more effective for a larger state space and a larger number of interactions; however we briefly highlight it here because of its widespread familiarity and because the information processing measures can be written in terms of “generalized” λ parameters (see Supplementary Materials). This means that λ ’s relation with the Wolfram classification is captured within the information features, implying that the information is minimally as predictive as features based on λ parameter(s).

Indeed, the predictive power $I(\lambda : c(R))/H(c(R)) \approx 0.175$, which is significantly lower than the information synergy alone which achieves 0.361 at $t = 1$. Moreover, as indicated by the black dots in Figure 2(a) the vast majority of information features have higher predictive power than 0.175; in fact only three single features have slightly lower power. It is not surprising yet reassuring that the information features outperform the Langton parameter.

3.1.4. Information Processing-Based Clustering. In the previous sections we showed how information features predict Wolfram’s behavioral classification. In this section we investigate the hierarchical clustering of rules induced by the information features in their own right. One important reason for studying this is the fundamental problem of undecidability of CA classifications based on long-term behavior characteristics [29, 30], such as for the Wolfram classification. In the best case this makes the classification difficult to compute; in the worst case the classification is impossible to compute, leading to questions about its meaning. In contrast, if local information features correlate strongly with long-term emergent behaviors, then a CA classification scheme based on information features is practically more feasible. In this section we visualize how the clustering overlaps with Wolfram’s classification.

Figure 3(a) shows a clustering made using information features evaluated where $t = 0$ is the randomized state. Interestingly, while the features have low predictability on the Wolfram class, the resulting clustering also does not overlap at all with Wolfram classification. We have to make an exception for rules 60, 90, 105, and 150 which are all chaotic rules and share the same information processing features.

FIGURE 3: Dendrograms displaying two clusterings of information processing features: (a) the information measures are computed between time $t = 0$ and $t = 1$, with no correlation between sites at time $t = 0$, and (b) when information processing features are evaluated numerically between time t and $t + 1$ once a steady state is reached. The color code is the same as in Figure 1 except that rules in class I are displayed in blue to distinguish them from class II.

Figure 3(b), on the other hand, displays the clustering for the case where features are not evaluated with respect to the randomized state but to the stationary distribution (i.e., $X^{t=0}$ is a randomly selected system state from the cyclic attractor). One reason for this is to ignore the initial transient phase; another reason is to make a step toward the financial application in the next subsection, which obviously does not start from random initial conditions. By “stationary” we mean that we simulate the CA dynamics long enough until the time-shifted information features no longer have a changing trend. Feature values can no longer be calculated exactly and thus are estimated numerically by sampling initial conditions and for size $N = 15$. In that case we find that the clustering has increased overlap with Wolfram’s classification. In particular, we can note that uniform rules cluster together and that chaotic and complex rules are all on the same large spray. However the agreement is far from perfect. For instance, the spray bearing chaotic and complex rules also bears periodic rules. Note also that rules 60, 90, 105, and 150 are indistinguishable when considered from the information processing viewpoint, even though they exhibit chaotic patterns that can be visually distinguished from each other. On the contrary, rules 106 and 154 are very close to each other and the pattern they exhibit indeed shows some similarities, but the former is complex while the latter is periodic.

Note that using this clustering scheme all rules converging to a uniform pattern, but one, are close to each other in the information features space. The remaining one, rule 168, has a transient regime which is essentially dominated by a translation of the initial state. This unexpected behavior is due to rare initial conditions (e.g., ...110110110...) that are present in our exact calculation with the same weight as all other

initial conditions but have a strong impact on the information processing measure. This translational regime can be found as well in rules 2 and 130, which are classified in the same subspray as rule 168. The similarity of any information feature (information transfer in this case) can thus lead to rules whose behavior differs in other respects to get classified similarly.

3.2. Detecting a Regime Shift Financial Time-Series. The results for the ECA models are promising but are under ideal and controlled conditions by construction: independent initial states, no external influences, and an enumerable list of dynamics. It is natural to ask whether the same formalism could eventually turn out valuable when analyzing real data of real complex systems, despite the fact that they do not obey such ideal conditions and cannot be controlled. A positive answer would add more urgency to further studying the idea of “information processing” systematically in models beyond ECA and toward real systems and data. A negative answer on the other hand could hint toward the idea being restricted to the specific realm of ECA and perhaps some similar simplistic models, but not being useful for studying real systems, in which case further systematic study would be less urgent. It is therefore important to find out whether there can be any hope of such a positive answer, which is the purpose of this section to demonstrate. More systematic studies remain nevertheless needed to understand how and why information processing features are predictive of emergent behaviors.

We focus on financial data because it is of high quality and available in large quantities. Also at least one large regime shift is known: the 2008 financial crisis, separating a precrisis and a postcrisis regime presumably by a “tipping point” phase of high systemic instability. We set the date of the financial

FIGURE 4: 200 time points showing the progression of the three information features memory (M), transfer (T), and integration (S) computed with a time delay of 1 day (similar to $t = 1$ for ECA). The color indicates the time difference with September 15, 2008 (big black dot), which we consider the starting point of the 2008 crisis, from dark blue (long before) to dark red (long after) and white at the crisis date. The data spans from January 1, 1999, to April 21, 2017; the large green dot is the last time point also present in the IRS data in 2011. In this information space we clearly observe signs of a two attractor regimes separated by a sudden regime shift. Mutual information is calculated using a sliding window of $w = 1400$ days; the 200 windows partially overlap and are placed uniformly over the dataset, where the first and last window include the first and last day of the dataset, respectively. The gray plus signs are the projections of the 3D points on the visible side faces for better visibility of the positions of the points.

crisis on September 15, 2008, which is the date of the Lehman Brothers bankruptcy.

We focus on two sequences of time-series: daily IRS rates of 14 and 15 maturities in the USD and EUR markets, and five consecutive daily FX closing exchange rates. We selected these datasets because the variables in each dataset can be considered to form a line graph similar to ECA rules, staying as close as possible to the previous analysis. Also, these two markets play a major role in today's global economy: the IRS market is the largest financial derivatives market, whereas the FX market is by far the largest trading market in terms of volume. Even though it remains yet unclear how exactly the crisis was driven by the different markets, we assume that we can at least measure a regime shift or a growing instability in each dataset.

At this point the information feature values are not (yet) understood in their own right, and the absolute value of the synergy feature is not meaningful because it is rescaled to avoid cointegration (see Methods). Nevertheless the information features do offer an alternative and potentially complementary way to parameterize the state space of the financial market, as opposed to using directly the observations themselves (interest or exchange rates in this case). For financial markets this is especially important because structures with predictive power quickly disappear from the system once observed, which is a rather unique property of financial markets. As Scheffer et al. [31] phrase it: "In

this field, the discovery of predictability quickly leads to its elimination, as profit can be made from it, thereby annihilating the pattern." Our proposed parameterization in terms of information features may not yet be exploited on a large-scale. This potentially means that the financial crisis may become detected or even anticipated when using standard model-free instability (or tipping point) indicators applied to the information features time-series, as opposed to using the original financial time-series data itself directly, which we explore in this section. In the end we propose a new model-free tipping point indicator for multivariate, nonstationary time-series applied to the main information features.

3.2.1. Foreign Exchange Market. In Figure 4 we show the 3-dimensional "information feature space" with the same axes as Figure 1. We observe remarkable separation of the precrisis and postcrisis periods which are well separated by a single transition trajectory immediately following the crisis date (black circle). Looking more closely, we also observe that early in the precrisis regime the information features traverse steadily but surely through the entire blue attractor (dark and medium blue dots). Directly preceding the regime shift the information features appear more clustered in one region in the lower part of the attractor (light blue dots), without a clear general direction. Soon after this "noisy stationary" phase there is evidently a clear direction again when it traverses from the blue to the red attractor. In the red attractor the

FIGURE 5: 200 time points showing the progression of the three information features memory (M), transfer (T), and synergy (S) computed with a time delay of 1 day (similar to $t = 1$ for ECA). The color indicates the time difference with September 15, 2008 (big black dot), which we take as the starting point of the 2008 crisis, from dark blue (long before) to dark red (long after) and white at the crisis date. The data spans more than twelve years: the EUR data from January 12, 1998, to August 12, 2011, and the USD data from April 29, 1999, to June 6, 2011. Mutual information is calculated using a sliding window of $w = 1400$ days; the 200 windows partially overlap and are placed uniformly over the dataset, where the first and last window include the first and last day of the dataset, respectively. The gray plus signs are the projections of the 3D points on the visible side faces for better visibility of the positions of the points.

system appears to steadily traverse through the attractor again; that is, it appears stationary on longer time scales but nonstationary on shorter time scales.

Interestingly, this behavior resembles to some extent the dynamics observed for the so-called tipping points [31] where a system is slowly pushed to an unstable point and then “over the hill” after which it progresses quickly “downhill” to the next attractor state. This is relevant because slow progressions to tipping points offer a potential for developing an early-warning signal.

3.2.2. Interest-Rate Swap Market. In Figure 5 we show the same feature space for the IRS markets in EUR and USD (EURIBOR and LIBOR). In short, interest-rate swaps consist of transactions exchanged between two agents such that their opposing risk exposures are effectively canceled. In contrast to the FX market, in IRS we observe the completely different scenario of steady nonstationary progressions of the information features during most of the duration of the dataset. One possible explanation is that these markets had not yet settled into an equilibrium, as they are relatively young markets (1986 and 1999 in their present form) continually influenced by market reforms and policy changes. A second possible explanation is that the contracts traded in this market are relatively long-term contracts, covering periods from a few months to a few decades, influencing subsequent traded contracts, whereas FX trades are instant and do not involve contracts.

Yet another possible but hypothetical explanation for this is that the IRS markets could have been (part of) a slow but steady driving factor in the global progression to the crisis, perhaps even building up a financial “bubble,” whereas the FX market may have been more exogenously forced toward their regime shift from one attractor to another. Indeed, the progression to the 2008 crisis is often explained by referring at least to substantial losses in fixed income and equity portfolios followed by the US subprime home

loan turmoil [32], suggesting at least a central role for the trade of risks concerning interest rates in USD. The exact sequence of events leading to the 2008 crisis is however still debated among financial experts. Our numerical analyses may nevertheless help to shed light on interpreting the relative roles of different markets.

In any case, in the EUR plot we observe that a steady and fast progression is following as well by a short “noisy stationary” period where there seems to be no general direction, after which a new and almost orthogonal direction is followed after the crisis. The evolution after the crisis is much more noisy, in the form of larger deviations around the general direction. In the USD we do not observe a brief stationary phase before the crisis, but we do observe larger deviations as well around the general directions, mostly sideways from the viewpoint of this plot. The market does contain two directional changes but these do not occur closely around the crisis point. We do not speculate here about their possible causes.

3.2.3. Potential Indicator for the Financial Crisis. All in all we observe in all three financial markets that the information features form a multidimensional trajectory which progresses in a (locally or globally) nonstationary manner. We also observe that around the crisis point the behavior appears characterized by increased variations around the same general trend (USD IRS) or by variations around a decreasing general trend. In this section we propose an instability indicator of this phase which can be applied to all three cases.

Several model-free (leading) indicators have been previously developed for time-series data in order to detect an (upcoming) tipping point for complex systems in general [31, 33, 34], such as for ecosystems, climate systems, and biological systems. By model-free we mean that the indicator is not developed especially for a particular dataset or domain and can easily be extended to other domains. These model-free instability indicators are computed from time-series

and include critical slowing down, variance, and correlation. However the financial system remains notoriously resilient to such analyses. One possible explanation for this is that the financial system has a tight feedback loop on its own state, and any known indicator with predictive power would soon be exploited and thus lead to behavioral changes in the market as long as it is present.

Regardless of the underlying reason, it has been shown that well-known model-free instability indicators hardly or not at all detect or predict the financial crisis. For instance, critical slowing down, variance, and correlations do not form a (leading) indicator in the same IRS data [11]. Babecký et al. [35] find that currency and banking crises are also hard to predict and resort to combinations of model-specific measures such as worsening government balances and falling central bank reserves. The only promising exception known to the present authors as a leading (early-warning) indicator is one based on specific network-topological features of inter-bank credit exposures (binary links), which shows detectable changes several years before the crisis [36]. Although it is specifically developed for interbank exposures, it could potentially be generalized to other complex systems as well, in cases where similar network data can be inferred. This remains nevertheless untested. All in all, the lack of progress inspired a number of renowned experts in complexity and financial systems [37] recently to call for an increased effort to bring new complexity methods to financial markets in order to better stabilize our financial markets.

Here we develop a new tipping point indicator for multi-dimensional data and test it on the sequence of information features of the three datasets. The currently well-known indicators are developed for univariate, stationary time-series and thus cannot be directly applied to (nonstationary) multivariate time-series, such as our information features.

We aim to generalize upon the idea of the variance indicator [34] which appears the most feasible candidate for multidimensional time-series. In contrast, computing critical slowing down involves computing correlations, which requires a large, combinatorially increasing amount of data as the number of dimensions grows. In short, the idea of the variance indicator is that prior to a tipping point the stability of the current attractor decreases, leading to larger variation in the system state, until the point where the stability is sufficiently low such that natural variation can “tip” the system over to a different attractor. This indicator is typically applied to one-dimensional system states, such as species abundance in ecosystems or carbon dioxide concentrations in climate, where the behavior in each attractor is assumed to be (locally) stationary.

A natural generalization of variance (or standard deviation) to higher dimensions is the average centroid distance: the average Euclidean distance of a set of points to their average (centroid). Since the centroid distance also increases when there is a directed general trend, which we wish to consider as natural behavior, we divide by the distance traversed by this general trend. The result in words is then the average centroid distance per unit length of trend. That is, for a sequence of Δt state vectors $\vec{x}_{t-\Delta t}, \vec{x}_{t-\Delta t+1}, \dots, \vec{x}_t$,

in our case information features, our indicator is defined as

$$\omega_{\Delta t}(t) = \frac{\langle \|\vec{x}_{t-i} - \langle \vec{x} \rangle\| \rangle_{0 \leq i \leq \Delta t}}{\|\vec{x}_t - \vec{x}_{t-\Delta t}\|}. \quad (11)$$

Here, Δt is the number of data points up to time t used in order to compute the indicator value at time t . Ideally, Δt should typically be as low as possible in order to provide an accurate local description of the system’s stability near time t , but not too low such that mostly noise effects are measured and/or the general trend cannot be distinguished effectively. To further filter out noise effects and study the indicator progression on different time scales, we use an averaging sliding window of g preceding indicator values to finally compute the indicator value at time t ; that is,

$$\omega_{\Delta t, g}(t) = \langle \omega_{\Delta t}(t') \rangle_{t-g < t' \leq t}. \quad (12)$$

Note that using these two subsequent sliding windows (Δt and g) is not equivalent to simply increasing Δt by g and then not averaging. To illustrate, imagine that the multivariate time-series forms a circle of $\Delta t + g$ points. Using a small Δt value relative to g will recover the fact that the circle is locally (almost) a straight line for each time t (low value in (11)), after which taking an average of g indicator values will result in a relatively low value in (12). In the extreme case of a straight line with uniformly spaced points, $\omega_{\Delta t, g}(t) = 1/4$. In contrast, increasing Δt by g and setting $g = 1$ means that the point x_t returns back adjacent to $x_{t-\Delta t}$ (so small denominator value in (11)) but a large average distance to the centroid (the radius of the circle). In the extreme case of $\Delta t \rightarrow \infty$, then $\omega_{\Delta t, g}(t) \rightarrow \infty$. This example makes it clear that Δt is preferably as low as possible to capture the short-term behavior; since this decreases the signal-to-noise ratio, we subsequently average over g most recent values.

Figure 6 shows the indicator values for all three datasets during roughly 7 (IRS) and 12 years (FX) around the crisis date. Strikingly, the short-term plots (Figure 6(a), $g = 10$) show that in all three cases there is a strong and sharp peak around the time of the crisis. For the FX market this peak precedes the crisis by almost one year; for the IRS market the peaks are just after the crisis (USD, 2-3 months) or well after the crisis (EUR, almost one year). Although there are also a few other peaks (EUR and FX), briefly discussed further below, it is reassuring that the indicator is capable of clearly detecting the financial crisis period around the Lehman Brothers collapse. The difference in timing is intriguing but not further studied here.

Note that the ω indicator values have an intuitive interpretation. A value of $1/4$ means that the multivariate time-series progresses in a perfectly straight line with uniform spacing. At another extreme, if the points are perfectly distributed in a symmetrically round cloud around the initial point, then ω tends to unity on average. If there is a directed trend but the orthogonal deviations are larger than the trend vector, then $\omega > 1$. In the Supplementary Materials we show indicator results for other parameter values and show that an alternative method of computing the nontrend variations

FIGURE 6: The proposed instability indicator (12) calculated for all three datasets. (a) is for small sliding window size ($g = 10$) and (b) for a long sliding window size ($g = 50$), showing the indicator both on short and long time scales. The indicator is computed from the sequence of 200 information feature vectors in Figures 4 and 5. The sliding window size (g) is illustrated by the gray bar in the top left of the right-hand panels. The first $g - 1$ indicator values are averages of fewer than g preceding values. One year corresponds to about 250 trade days. Note that the x -axes left and right are different. $\Delta t = 10$.

indicator, by projecting all points to the “trend” vector $\vec{x}_t - \vec{x}_{t-\Delta t}$ and computing the average perpendicular (projection) distance, gives very similar results.

It is common to study instability indicators at a larger time scale in order to detect or even predict the largest events, ignoring (smoothing out) smaller events. In particular the hope is to find a *leading* indicator which could be used to anticipate the 2008 onset of recession. We show the same indicator but now averaged over $g = 50$ values in Figure 6(b). Remarkably, all three datasets show a discernible, long-term steady growth in the instability indicator leading up and through the crisis date. For the EUR and FX curves this growth starts around two years before the crisis; for

the USD curve the growth starts at the start of the curve. Although here the initial peak in the EUR curve appears to even outweigh the crisis-time peak, we must note here that this peak is subject to less smoothing as there are fewer than g values to the left available for averaging (compare with the sliding window size depicted in gray); with further averaging all other peaks will continue to decrease in height (cf. Figure 6(a)), whereas this initial peak will remain roughly at the same value for this reason.

We will now discuss two significant additional strong peaks observable in the indicator curves: an initial peak in EUR (around August-September 2004) and a late peak in FX (mid-2016). We caution that it is hardly scientific to

reason back from observed peaks toward potential underlying causes, especially for continually turbulent systems such as the financial markets where events are easy to find. Nevertheless it is important to evaluate whether the additional peaks at least *potentially* could indicate substantial systemic instabilities, or whether they appear likely to be false positives.

For the EUR initial peak we refer to ECB's *Euro Money Market Study*, May 2004 report. We find that this report has indeed an exceptionally negative sentiment compared to other years, speaking of "declining and historically low interest rates," an inverted yield curve, "high geopolitical tensions in the Middle East and the associated turbulence in oil prices and financial markets," and "growing pessimism with regard to economic growth in the euro area." Also: "The ECB introduced some changes to its operational framework which came into effect starting from the main refinancing operation conducted on 9 March 2004." In contrast, the subsequent report (2006) is already much more optimistic: "After two years of slow growth, the aggregated turnover of the euro money market expanded strongly in the second quarter of 2006. Activity increased across all money market segments except in cross-currency swaps, which remained fairly stable." We deem it at least plausible that the initial EUR indicator peak, which has about half the height of the after-crisis peak, is a true positive and detects indeed a period of increased systemic instability or stress.

For the more recent FX peak across 2016 we must refer to news articles such as in Financial Times [38, 39]. Firstly there were substantial and largely unanticipated political shifts, including Brexit (dropping Sterling by over 20%) and the election of Trump as US President. At the same time, articles mention fears about China's economic growth slowing down. Lastly, as interest rates affect associated currencies: "By August, global [bond yield] benchmarks were at all-time lows, led by the 10-year gilt yielding a paltry 0.51 per cent, while Switzerland's entire bond market briefly traded below zero. [...] The universe of negative yielding debt had swollen to \$13.4tn." For example, earlier in the year (January 29), the Bank of Japan unexpectedly started to take their interest rates into the negative for the first time, affecting the Yen. Questions toward another recession are mentioned, although also discarded. All in all, we deem it at least plausible that the FX market's indicator peak in 2016 could be caused indeed by systemic instability and stress, that is, a true positive.

All in all we deem the proposed "normalized centroid distance" instability indicator as a high potential candidate for multivariate, nonstationary time-series. Secondly we argue that parameterizing a market state in terms of information features instead of the original observations (interest or exchange rates) is useful and enables detecting growing systemic instability. However we must caution that our financial data only contains one large-scale onset of recession (2008), so it is difficult to provide conclusive validation that such events are detected reliably by the proposed indicator. Future work may include applying the indicator to different simulated systems which can be driven toward a large-scale regime shift.

4. Discussion

Our working assumption is that dynamical systems inherently process information. Our leading hypothesis is that the way that information is locally processed determines the global emergent behavior. In this article we propose a way to quantitatively characterize the notion of information processing and assess its predictive power of the Wolfram classification of the eventual emergent behavior of ECA. We also make a "leap of faith" to real (financial) time-series data and find that transforming the original time-series to an information features time-series enables detection of the 2008 financial crisis by a simple leading indicator. Since it is known that the original data does not permit such detection, this suggests that novel insights may be gained even in real data of complex systems, despite not obeying the ideal conditions of our ECA approach. This warrants a further systemic study into this notion of information processing in different types of models and eventually datasets.

Our formalization builds upon Shannon's information theory, which means that we consider an ensemble of state trajectories rather than a single trajectory. That is, we do not quantify the information processing that occurs during a particular, single sequence of system states (attempts to this end are followed by Lizier et al. [40]). Rather, we consider the ensemble of all possible state sequences along with their probabilities. One way to interpret this is that we quantify the "expected" information processing averaged over multiple trajectories. Another way to interpret it is that we characterize a dynamical *model* in its totality, rather than a particular symbolic sequence of states of a model. Our reasoning is that if (almost) every state trajectory of a model (such as a CA rule) leads to a particular emergent behavior (such as chaotic or complex patterning), then we would argue that the emergent behavior is a function of the ensemble dynamics of the model.

This seems at odds with computing information features from real time-series, which are measurements of a single trajectory of a system. We resolve this issue by assuming "local stationarity." This assumption is common in time-series analysis and used (implicitly or explicitly) in most "sliding window" approaches and moving statistic estimations, among others. In other words, we assume that the rate of sampling data points is significantly faster than the rate at which the underlying statistical property changes, which in our case are the information features. The consequence is that a finite number of consecutive data points can be used to estimate the probability distribution of the system at the corresponding time, which in turn enables estimating mutual information quantities.

Our first intriguing result from the ECA analysis is that fewer information features capture more of the relevant dynamical behavior, as time t progresses away from a randomized system state. One potential explanation is the processing of correlated states or, equivalently, of overlapping information. Namely, to reach at $t = 1$ each cell operates exclusively on uncorrelated inputs, so the resulting state distribution is a direct product of the state transition table, irrespective of how cells are connected to each other.

Neighboring cell states at $t = 1$ are correlated due to overlapping neighborhoods in this network of connections. Consequently, at time $t = 2$ and beyond, the inputs to each cell have become correlated in a manner dictated by the interaction topology. The way in which an ECA rule subsequently deals with these correlations is evidently an important characteristic. In other words, two ECA rules may have exactly the same information features for $t = 1$ but different features for $t = 2$ which must be due to a different way of handling correlated states.

This result leads us to hypothesize an *information locality* concept. That is, the information features at $t = 1$ of each cell do not yet characterize the interaction topology (by the correlations it induces). In other words, all interaction topologies where each cell has 3 neighbors are indistinguishable at $t = 1$. This suggests that the “missing” predictive power at $t = 1$ is a measure of the relevance of the interaction topology. In the case of ECA this quantity is roughly half: $1 - 0.49 = 0.51$, where 0.49 is the maximal predictive power at $t = 1$. For the sake of illustration, suppose that the eventual behavior exhibited by a class of systems depends crucially and only on the number of neighbors at distance 5. In this case we expect that the predictive power of the information features does not reach 1.0 until at least $t \geq 5$, since otherwise the cell states have not had the opportunity to be causally influenced by the local network structure at distance 5. If no further network effects at larger distances play a role, then we expect the predictive power to reach exactly 1.0 at $t = 5$. For ECA we find 1.0 predictive power already at $t = 3$, suggesting that there are no nonlocal network features which play a role in the dynamics, which is indeed true for ECA (all ECA are uniform, infinite-length line graphs). We hypothesize that the distance, that is, *time step* t at which the predictive power no longer increases, is a measure of the “locality” of topological characteristics that are relevant for the emergent behavior. However we leave it for future work to fully investigate this concept.

Our second intriguing result is that the most predictive information feature is invariably *synergy*. In each time step it accounts for the vast majority of the total predictive power (75%, 92%, and 96%, resp.). This is the feature that we would consider to actually capture the “processing” or integration of information, rather than the memory and transfer features which capture the simple “copying” of information. Indeed, the cube of Figure 1 suggests that the interesting behaviors (chaotic and complex) are associated with high synergy, low memory, and low transfer. In this extreme we find rule 60 (the XOR rule) and similar rules which are all chaotic rules. For complex behavior nonzero but low memory and transfer appear to be necessary ingredients.

The good separation of the dynamic behavioral classes in the ECA models using only a few information features ultimately leads to the question whether the same can be done for real systems based on real data. This is arguably a large step and certainly more rigorous research should be done using intermediate models of increasing complexity and for different classifications of dynamical behavior. On the other hand, if promising results could be obtained from real data using a small set of information features, then this would

add more urgency to such future research, even if the role of information processing features in systemic behavior is not yet fully understood. This is the purpose of our application to financial data. Financial data is of high quality and available in large quantities and at least one large regime shift is known, namely, the 2008 financial crisis. We stay as close as possible to ECA by selecting two datasets in which the dynamical variables could be interpreted to form a line graph. Although of course many forces which act on the markets are not contained in these datasets, the information features may still be able to detect changes in the correlations over time, despite not knowing what is the root cause of these changes. In fact, a primary driver behind our approach is indeed the abstraction of physical or mechanistic details while still capturing the emergence of different types of behaviors. We consider our results in the financial application promising enough to warrant further study into information processing features in complex system models and other real datasets. Our results suggest tipping point behavior for the FX and EUR IRS markets and a possible driving role for the USD IRS market.

All in all we conclude that the presented information processing concept appears indeed to be a promising direction for studying how dynamical systems generate emergent behaviors. In this paper we present initial results which support this. Further research may identify concrete links between information features and various types of emergent behaviors, as well as the relative impact of the interaction topology. Our lack of understanding of emergent behaviors is exhibited by the ECA model: it is arguably the simplest dynamical model possible, and the choice of local dynamics (rule) and initial conditions fully determine the emergent behavior that is eventually generated. Nevertheless even in this case no theory exists that predicts the latter from the former. The information processing concept may eventually lead to a framework for studying how correlations behave in dynamical systems and how this leads to different emergent behaviors.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

Peter M. A. Sloot and Rick Quax acknowledge the financial support of the Future and Emerging Technologies (FET) Programme within Seventh Framework Programme (FP7) for Research of the European Commission, under the FET-Proactive grant agreement TOPDRIM, no. FP7-ICT-318121. All authors also acknowledge the financial support of the Future and Emerging Technologies (FET) Programme within Seventh Framework Programme (FP7) for Research of the European Commission, under the FET-Proactive grant agreement Sophocles, no. FP7-ICT-317534. Peter M. A. Sloot acknowledges the support of the Russian Scientific Foundation, Project no. 14-21-00137.

Supplementary Materials

Figure 1: total information evaluated for rule 110 over 50 time steps, starting from an uncorrelated initial state. Figure 2: memory information evaluated for rule 110 over 50 time steps, starting from an uncorrelated initial state. Figure 3: left-transfer information evaluated for rule 110 over 50 time steps, starting from an uncorrelated initial state. Figure 4: right-transfer information evaluated for rule 110 over 50 time steps, starting from an uncorrelated initial state. Figure 5: 200 time points showing the progression of the three information features memory (M), transfer (T), and synergy (S) computed with a time delay of 1 day (similar to $t = 1$ for ECA). The color indicates the time difference with September 15, 2008 (big black dot), which we consider the starting point of the 2008 crisis, from dark blue (long before) to dark red (long after) and white at the crisis date. The data spans from January 1, 1999, to April 21, 2017; the large green dot is the last time point also present in the IRS data in 2011. Mutual information is calculated using a sliding window size of $w = 1000$ days; the 200 windows partially overlap and are placed uniformly over the dataset, where the first and last window include the first and last day of the dataset, respectively. Figure 6: 200 time points showing the progression of the three information features memory (M), transfer (T), and synergy (S) computed with a time delay of 1 day (similar to $t = 1$ for ECA). The color indicates the time difference with September 15, 2008 (big black dot), which we consider the starting point of the 2008 crisis, from dark blue (long before) to dark red (long after) and white at the crisis date. The data spans more than twelve years: the EUR data from January 12, 1998, to August 12, 2011, and the USD data from April 29, 1999, to June 6, 2011. Mutual information is calculated using a sliding window of $w = 1000$ days; the 200 windows partially overlap and are placed uniformly over the dataset, where the first and last window include the first and last day of the dataset, respectively. Figure 7: 200 time points showing the progression of the three information features memory (M), transfer (T), and synergy (S) computed with a time delay of 1 day (similar to $t = 1$ for ECA). The color indicates the time difference with September 15, 2008 (big black dot), which we consider the starting point of the 2008 crisis, from dark blue (long before) to dark red (long after) and white at the crisis date. The data spans from January 1, 1999, to April 21, 2017; the large green dot is the last time point also present in the IRS data in 2011. Mutual information is calculated using a sliding window of $w = 2000$ days; the 200 windows partially overlap and are placed uniformly over the dataset, where the first and last window include the first and last day of the dataset, respectively. Figure 8: 200 time points showing the progression of the three information features memory (M), transfer (T), and synergy (S) computed with a time delay of 1 day (similar to $t = 1$ for ECA). The color indicates the time difference with September 15, 2008 (big black dot), which we consider the starting point of the 2008 crisis, from dark blue (long before) to dark red (long after) and white at the crisis date. The data spans more than twelve years: the EUR data from January 12, 1998, to August 12, 2011, and the USD data from April 29, 1999, to June 6, 2011. Mutual information is calculated using a sliding

window of $w = 2000$ days; the 200 windows partially overlap and are placed uniformly over the dataset, where the first and last window include the first and last day of the dataset, respectively. Figure 9: indicator curves as in main text's Figure 6 but using perpendicular distances to the trend vector as alternative to (11). (*Supplementary Materials*)

References

- [1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, vol. 6, John Wiley & Sons, 1991.
- [2] C. G. Langton, "Computation at the edge of chaos: phase transitions and emergent computation," *Physica D: Nonlinear Phenomena*, vol. 42, no. 1–3, pp. 12–37, 1990.
- [3] P. Grassberger, "Toward a quantitative theory of self-generated complexity," *International Journal of Theoretical Physics*, vol. 25, no. 9, pp. 907–938, 1986.
- [4] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya, "Information modification and particle collisions in distributed computation," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 20, no. 3, Article ID 037109, 2010.
- [5] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya, "The information dynamics of phase transitions in random boolean networks," in *Proceedings of the 11th International Conference on the Simulation and Synthesis of Living Systems: Artificial Life XI, ALIFE 2008*, pp. 374–381, 2008.
- [6] R. D. Beer and P. L. Williams, "Information processing and dynamics in minimally cognitive agents," *Cognitive Science*, 2014.
- [7] E. J. Izquierdo, P. L. Williams, and R. D. Beer, "Information flow through a model of the *C. elegans* klinotaxis circuit," <https://arxiv.org/ftp/arxiv/papers/1603/1603.03552.pdf>.
- [8] Y. Bar-Yam, D. Harmon, and Y. Bar-Yam, "Computationally tractable pairwise complexity profile," *Complexity*, vol. 18, no. 5, pp. 20–27, 2013.
- [9] B. Allen, B. C. Stacey, and Y. Bar-Yam, "An Information-Theoretic Formalism for Multiscale Structure in Complex Systems," <https://arxiv.org/abs/1409.4708v1>.
- [10] R. Quax, A. Apolloni, and P. M. A. Slood, "The diminishing role of hubs in dynamical processes on complex networks," *Journal of the Royal Society Interface*, vol. 10, no. 88, 2013.
- [11] R. Quax, D. Kandhai, and P. M. A. Slood, "Information dissipation as an early-warning signal for the Lehman Brothers collapse in financial time series," *Scientific Reports*, vol. 3, article no. 1898, 2013.
- [12] K. Lindgren, "An information-theoretic perspective on coarse-graining, including the transition from micro to macro," *Entropy*, vol. 17, no. 5, pp. 3332–3351, 2015.
- [13] R. G. James, C. J. Ellison, and J. P. Crutchfield, "Anatomy of a bit: Information in a time series observation," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 21, no. 3, Article ID 037109, 2011.
- [14] P. L. Williams and R. D. Beer, "Nonnegative decomposition of multivariate information," <https://arxiv.org/abs/1004.2515v1>.
- [15] E. Olbrich, N. Bertschinger, and J. Rauh, "Information decomposition and synergy," *Entropy*, vol. 17, no. 5, pp. 3501–3517, 2015.
- [16] R. Quax, O. Har-Shemesh, and P. M. A. Slood, "Quantifying synergistic information using intermediate stochastic variables," *Entropy*, vol. 19, no. 2, article no. 85, 2017.
- [17] G. Chliamovitch, B. Chopard, and L. Velasquez, "Assessing complexity by means of maximum entropy models," <https://arxiv.org/abs/1408.0368>.

- [18] V. Griffith, E. K. P. Chong, R. G. James, C. J. Ellison, and J. P. Crutchfield, "Intersection information based on common randomness," *Entropy*, vol. 16, no. 4, pp. 1985–2000, 2014.
- [19] V. Griffith and T. Ho, "Quantifying redundant information in predicting a target random variable," *Entropy*, vol. 17, no. 7, pp. 4644–4653, 2015.
- [20] S. Wolfram, *A New Kind of Science*, Wolfram Media, Champaign, Ill, USA, 2002.
- [21] R. D. Beer and P. L. Williams, "Information processing and dynamics in minimally cognitive agents," *Cognitive Science*, vol. 39, no. 1, pp. 1–38, 2015.
- [22] N. Timme, W. Alford, B. Flecker, and J. M. Beggs, "Synergy, redundancy, and multivariate information measures: an experimentalist's perspective," *Journal of Computational Neuroscience*, vol. 36, no. 2, pp. 119–140, 2014.
- [23] E. Schneidman, W. Bialek, and M. J. Berry II, "Synergy, Redundancy, and Independence in Population Codes," *The Journal of Neuroscience*, vol. 23, no. 37, pp. 11539–11553, 2003.
- [24] Wolfram—Alpha: Computational Knowledge Engine, 2015. <http://www.wolframalpha.com/>.
- [25] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 5, no. 4, pp. 537–550, 1994.
- [26] T. W. S. Chow and D. Huang, "Estimating optimal feature subsets using efficient estimation of high-dimensional mutual information," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 16, no. 1, pp. 213–224, 2005.
- [27] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, no. 6, article 066138, 2004.
- [28] <http://www.global-view.com/forex-trading-tools/forex-history/>.
- [29] I. Culik and S. Yu, "Undecidability of CA classification schemes," *Complex Systems*, vol. 2, no. 2, pp. 177–190, 1988.
- [30] K. Sutner, "Computational classification of cellular automata," *International Journal of General Systems*, vol. 41, no. 6, pp. 595–607, 2012.
- [31] M. Scheffer, J. Bascompte, W. A. Brock et al., "Early-warning signals for critical transitions," *Nature*, vol. 461, no. 7260, pp. 53–59, 2009.
- [32] M. Melvin and M. P. Taylor, "The crisis in the foreign exchange market," *Journal of International Money and Finance*, vol. 28, no. 8, pp. 1317–1330, The Global Financial Crisis: Causes, Threats and Opportunities, 2009.
- [33] V. Dakos, M. Scheffer, E. H. Van Nes, V. Brovkin, V. Petoukhov, and H. Held, "Slowing down as an early warning signal for abrupt climate change," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 38, pp. 14308–14312, 2008.
- [34] V. Dakos, E. H. van Nes, R. Donangelo, H. Fort, and M. Scheffer, "Spatial correlation as leading indicator of catastrophic shifts," *Theoretical Ecology*, vol. 3, no. 3, pp. 163–174, 2010.
- [35] J. Babecký, T. Havránek, J. Matějů, M. Rusnák, K. Šmídková, and B. Vašíček, "Banking, debt, and currency crises in developed countries: Stylized facts and early warning indicators," *Journal of Financial Stability*, vol. 15, pp. 1–17, 2014.
- [36] T. Squartini, I. Van Lelyveld, and D. Garlaschelli, "Early-warning signals of topological collapse in interbank networks," *Scientific Reports*, vol. 3, article no. 3357, 2013.
- [37] S. Battiston, J. D. Farmer, A. Flache et al., "Complexity theory and financial regulation: Economic policy needs interdisciplinary network analysis and behavioral modeling," *Science*, vol. 351, no. 6275, pp. 818–819, 2016.
- [38] The big events that shook financial markets in 2016. <https://www.ft.com/content/6d24125c-c066-11e6-9bca-2b93a6856354>.
- [39] Major Economic Events 2016: What Moved the Markets This Year? <https://www.orbex.com/blog/2016/12/major-economic-events-2016-moved-markets-year-infographics>.
- [40] J. T. Lizier, *The Local Information Dynamics of Distributed Computation in Complex Systems*, Springer, 2012.

Submit your manuscripts at
www.hindawi.com